Introduction to Unix, hands-on exercises

Course: Bioinformatics 2, IT and health.
Center for Biological Sequence Analysis, Technical University of Denmark.

Preview of Unix commands

Login to the system using the username and password combination provided to you. You will come to your home directory.

Enter the following commands and interpret the output.

```
echo hello world
date
who am i
who
man echo
clear
echo 2+4 | bc -l
ls
mkdir parent1
cd parent1
cd ..
```

Exercise Part 1

Login to the system using the usernames and password provided to you.
You will come to your home directory.

1. Create a directory called *work*.

2. Change directory to *work*.

3. Create a file called *hello.txt.*

On the command prompt, type *cat > hello.txt*

Write in some content you want. When finished, press Ctrl+D to close the file.

4. Check whether the file creation was successful. List the directory contents.

5. What is the file size? Try *ls* with appropriate options.

6. Check the contents of the file you just created. This can be done in two ways. Try both.

7. Move back to the parent directory (work).

8. Create another directory *play*.

9. Change to the directory *play* and create three sub-directories:

child1, *child2* and *child3*.

Can you create all three directories using a single command?

10. List the contents of the current directory (*play*). What do you see?

11. Change the directory to *child1*.

12. Create a file, *childfile*; type in some content and close.

13. Change to the home directory.

14. Copy file *work/hello.txt* to *child1* sub-directory without changing to any directories [stay in your home directory]. List the contents of *work.* How many files do you see?

15. Rename your *work/hello.txt t*o *myFile.txt* [stay in your home directory]. List the contents of *work.* How many files do you see?

16. Copy *Random_text.txt* from */home/people/tejal/public_html/courses/* to your *work* directory.

17. Delete directory *work*.

Exercise Part 2

1. Copy all files from */home/people/tejal/public_html/courses/* to your home directory.

2. List files with names starting with "gene".

3. There are three filenames starting with geneIDpart (*geneIDpart1.txt, geneIDpart2.txt, geneIDpart23.txt* and *geneIDpartitiion.txt*). How would you list files *geneIDpart1.txt* and *geneIDpart2.txt* but not *geneIDpart23.txt* ?

4. If you did not know the extension of the these files (.txt) how would you select *geneIDpart2.txt* but not *geneIDpart23.txt* ?

5. How many lines do the files starting with “pau” have ? (Hint: *wc* command with wildcards).

6. Print first 10 lines of *geneIDs.txt* (Hint: *head* command).

7. Print last 3 lines of *geneIDs.txt*.

8. How many times does "ENST0000015" appear in *geneIDs.txt* ?

9. How many times does "ENSG0000015" appear in *geneIDs.txt* ?

10. Can you find genes and/or transcript IDs with accession number containing "0000015151"? [regular expression].

11. From the file *Random_text.txt*, identify lines that do not start with "HGNC".

12. From file *Random_text.txt,* print line numbers of empty lines [those with nothing but one or more spaces].

13. In *Random_text.txt* find the lines that contain "*ST13P*". These are pseudogenes (hence the P) of different variety. Can you find how many *ST13* pseudogenes with numbers 11 to 19. [Hint: use grep regular expression].

14. Take a look at the "*evalSet*" file in your directory. It contains peptide sequences and numbers written in the Scandinavian number writing style. They are in fact, decimal points. Can you only extract the first column in this file? Can you direct the output of your command to a file ? [Hint: You need to specify the delimiter.]

15. You saw that those numbers in *evalSet* file. Commas in those numbers should show decimal points instead. How would you convert comma to decimal-point ?

16. You also saw that the columns in *evalSet* are separated by a semicolon (;), how would you convert form semi-colon to tab? Hint: tab is written as "\t" in computer understandable format.

17. So far you have only printed the files after conversion. You have converted formats of files without actually changing them. Can you combine the work done in number 12 and number 13 (that is, comma to decimal point and semicolon to tab) in a single command? It still won't change the original file or will the output get stored anywhere. Can you redirect the output of the command to a file named "*eval_66_log50k*" ?

18. Can you print *five* peptide sequences with highest score from *eval_66_log50k* (second column) ?